Supporting Information

Hampton *et al.* 10.1073/pnas.0711099105

SI Methods

Computational Models. ***Reinforcement learning.*** Reinforcement learning (RL) is concerned with learning predictions of the future reward that will be obtained from being in a particular state of the world or performing a particular action. In this paper we use a simple RL model in which action values are updated via a Rescorla-Wagner (RW) rule (1). On a trial t in which action a is selected, the value of action a is updated via a prediction error δ:

$$V^a_{t+1} = V^a_t + \eta\delta_t, \quad [1]$$

where η is the learning rate. The prediction error δ_t is calculated by comparing the actual reward received R_t after choosing action a with the expected reward for that action:

$$\delta_t = R_t - V^a_t. \quad [2]$$

When choosing between two different states (a and b), the model compares the expected values to select which will give it the most reward in the future. The probability of choosing action a is

$$p^a = f(V^a - V^b), \quad [3]$$

where $f(z) = 1/1 + e^{-\beta z}$ is the Luce choice rule (2) or logistic sigmoid, and β reflects the degree of stochasticity in making the choice (i.e., the exploration/exploitation parameter).

Fictitious play. In game theory, a first order fictitious play model (3) is one in which a player infers the probability that the opponent will choose one action or another, and then decides so as to maximize the action's consequent expected reward. The opponent's probability p^* of choosing an action a' is dynamically inferred by tracking the history of actions the opponent makes:

$$p^*_{t+1} = p^*_t + \eta\delta^p_t, \quad [4]$$

where $\delta^p_t = P_t - p^*_t$ is the prediction error between the opponent's expected action p^* and whether the opponent chose action a' at time t ($P = 1$), or chose another action ($P = 0$). Given the opponent's action probabilities p^*, the expected value for each of the player's actions can be calculated using the payoff matrix of the game. A stochastic choice probability can then be calculated using Eq. **3**. For the inspection game described in this paper, this can be summarized as follows: calling p the probability that the employee will work, and q the probability that the employer will not inspect, and using the payoff matrix of the game (Table 1—in the following formulations, payoffs were express in 25 cent units for convenience), the decision of each player is

$$p = f(2 - 4q^*) \quad [5]$$
$$q = f(5p^* - 1),$$

where q^* and p^* are the inferred probabilities of the opponent's actions estimated using Eq. **4**.

An equivalent formulation of fictitious play is one in which the values of actions are learned directly as in reinforcement models, instead of tracking the opponent player's action probability. For this, not only the value of the chosen action is updated with the reward that was received (as in Eq. **1**), but also all other actions are penalized proportional to their foregone rewards (4, 5). Either approach posits knowledge of the structure of the game to update the variable estimates and arrive at a correct expected value for the actions of each player.

Experience Weighted Attraction (EWA): The EWA learing rule we used here is a combination of Reinforcement Learning and Fictitious play. It updates the value of a choice such that:

$$V^a_{t+1} = (1 - \eta)V^a_t + \eta(\delta\cdot R_1 + (1 - \delta)\cdot R_2),$$

where R_1 is the reward obtained had action a been chosen (Fictitious learning), and R_2 is the reward given that action a was chosen − zero otherwise (Reinforcement Learning). Some variants of EWA also involve an additional parameter which modulates the rate of learning at different points in the game, such that it can be faster at the beginning of a game, and then become slower as subjects settle into a strategy towards the end. In this study we assumed cnstant learning throughout the game, so did not include this additional parameter.

Influencing the Opponent. How much does a player's next decision change given the action of the opponent? Replacing the update of the inferred opponent's strategy (Eq. 4) in a player's decision (Eq. **5**), and Taylor expanding (around $\eta = 0$),

$$\Delta p \approx -\eta 4\beta p_t(1 - p_t)(Q_t - q^*_t) \quad [6]$$
$$\Delta q \approx +\eta 5\beta q_t(1 - q_t)(P_t - p^*_t).$$

The sign difference in both terms is determined by the competitive structure of the game; namely, that the employer wants to inspect when the employee shirks, while the employee wants to shirk when the employer does not inspect. A player can obtain a more accurate inference of the opponent's action strategy by incorporating the influence his/her own action has on the opponent. Thus, at the end of each trial both players update the estimates of their opponent such that

$$p^*_{t+1} = p^*_t + \eta_1(P_t - p^*_t) - \eta_2 4\beta p^*_t(1 - p^*_t)(Q_t - q^{**}_t) \quad [7]$$
$$q^*_{t+1} = q^*_t + \eta_1(Q_t - q^*_t) + \eta_2 5\beta q^*_t(1 - q^*_t)(P_t - p^{**}_t);$$

where q^{**} and p^{**} are the inferred probabilities that the opponent has of the player itself (second-order beliefs). Thus, this gives two clear signals: the prediction error as the first term and the influence update as the second term. The influence update, or how much a player influences his/her opponent, is proportional to the difference between the action a player took and what the opponent thought was the player's strategy. These second order beliefs can be inferred by the player directly from the inferred opponent's strategy by inverting Eq. **5**.

$$p^{**}_t = \frac{4}{5} - \frac{1}{5\beta}log\left(\frac{1 - q^*_t}{q^*_t}\right) \quad [8]$$
$$q^{**}_t = \frac{1}{2} + \frac{1}{4\beta}log\left(\frac{1 - p^*_t}{p_t}\right).$$

Behavioral Data Analysis. The RL, Fictitious play, and Influence model decision probabilities p(action a or b)—the probability a certain action would be taken predicted by the model (Eq. **3**)—were fitted against the behavioral data y(action a or b)—the actual behavioral choice made by the subject. The parameters of each model were fitted against all subjects responses by maximizing the logistic log likelihood of the model predictions:

$$\log L = \frac{1}{N_{subjects}} \sum_{subjects} \left(\frac{1}{N_{trials}} \sum_{trials} y_a \log p_a + y_b \log p_b \right)^{Employer} + \frac{1}{N_{subjects}} \sum_{subjects} \left(\frac{1}{N_{trials}} \sum_{trials} y_a \log p_a + y_b \log p_b \right)^{Employee} \quad [9]$$

with one set of parameters modeling the employer role, and another the employee role. We used the multivariate constrained minimization function (fmincon) of the Optimization Toolbox 2.2 in MATLAB 6.5 (MathWorks) to estimate the model parameters given the log likelihood defined above.

The fitted parameters for each model are shown in Table S1.

fMRI Model Comparison: Voxel-Based Analysis. To compare the explanatory power of signals predicted by two competing models, we fit both models simultaneously to the brain BOLD signals (Fig. 2*B*) and then test for significance of a particular regressor (in this case, a random effects *t* test of the Influence model's expected reward signal). When testing a particular regressor for significance, SPM uses the extra sum of squares test by comparing the variance explained by a full model containing this regressor to a partial model not containing that regressor, e.g., (SS(Influence + RL) − SS(RL))/MSE; where MSE = residual error (6). In Fig. 2*B*, we report the additional variance explained by value signals from the Influence model above and beyond that explained by such signals from the RL model. It should be noted that because the free parameters for each model were fit on the behavioral data and not on the imaging data, the model comparison procedure on the imaging data are not affected by the different number of parameters in the model-fits to the behavioral data. Thus, there is no need to correct for different numbers of free parameters in the models (using for instance AIC methods) when performing the model-comparison procedure on the fMRI data.

fMRI Model Comparison: ROI Analysis. A second approach consists of fitting both models separately to brain activity in a region of interest, and then comparing their regression coefficients to determine which model provides a better account of neural activity in this region. To test how well the expected reward signals from each model predicted BOLD activity in mPFC we extracted a deconvoluted time series at the time of trial onset from all voxels in an 8 mm ROI centered on co-ordinates for mPFC derived from the metaanalysis of Frith and Frith (7). The process for extracting deconvoluted time series is explained in the section on inter-region correlation analysis (see below). We mean corrected and normalized the variance of the expected reward regressors from both models, and for each subject, fit each model to the deconvoluted time series. The regression coefficient of a particular model indicates how well the expected reward signal from that model correlates with BOLD activity at the time of trial onset. Thus, by comparing the Influence and RL subject-wise regression coefficients we can determine whether the Influence or the RL expected reward signals are a better predictor of brain activity in this region of interest. This comparison was done using a pair-wise *t* test across subjects.

Interregion Correlation Analysis. To compute inter-region correlations as reported in Fig. 4, a representative time series was obtained for each region by extracting and averaging BOLD activity from 10-mm spheres centered in the group peak for expected value in mPFC (−3, 63, 15 mm; see Fig. 2*A*); in the group peak for reward prediction errors in ventral striatum (9, 6, −18 mm, and −9, 9, −18 mm; see Fig. S3); and in the group peak for influence error signals in STS (−57, −54, 0 mm, and 60, −54, 9 mm; see Fig. 3*A*). A general linear model with one regressor for each trial (totaling 100 regressors) was then created by convoluting a canonical hemodynamic response function with a stick function centered at the time of trial onset. This model was then fitted to the time series from each region of interest. The regression coefficients thus represent the deconvoluted neural activity for each trial of the game at the time of trial onset. The deconvoluted activity from each region of interest was then used to calculate the correlation between regions of interest at that time point. The deconvolution and correlation process was then repeated for each different time point within a trial to plot the change in correlation between regions with respect to time within a trial for all subjects scanned in this study.

We also investigated the degree to which the time series extracted from STS and ventral striatum, as described above, were a significant predictor of the time series extracted from mPFC. In particular, we were interested in whether a linear model containing signals from STS and ventral striatum provided more information when predicting activity in mPFC, than did linear models which only contained signals from either STS or ventral striatum. For example, to determine whether adding STS as a regressor to a model that already contains ventral striatum as a regressor was a better predictor of activity in mPFC than the model with ventral striatum alone, we compared the likelihood of the complete model ($L_{\mathrm{STS,vStriatum}}$) with the likelihood of the incomplete model ($L_{\mathrm{vStriatum}}$). The difference in log likelihoods ($\Lambda = 2 \log L_{\mathrm{vStriatum}} - 2 \log L_{\mathrm{STS,vStriatum}}$) follows a χ^2 distribution with one degree of freedom. Thus, for each subject scanned (and each game session), the probability of whether adding information from STS provided statistically significant information was calculated. A random effects statistic across subjects was then calculated by converting these individual *p*-values to z-scores, and then performing a *t* test across subjects.

1. Rescorla RA, Wagner AR (1972) in *Classical conditioning II: Current Research and Theory*, eds Black AH, Prokasy WF (Appleton-Centrury-Crofts, New York), pp 64–99.
2. Luce DR (2003) *Response Times* (Oxford Univ Press, New York).
3. Fudenberg D, Levine DK (1998) *The Theory of Learning in Games* (MIT Press, Cambridge, MA).
4. McClure SM, Berns GS, Montague PR (2003) Temporal prediction errors in a passive learning task activate human striatum. *Neuron* 38:339–346.
5. Lohrenz T, McCabe K, Camerer CF, Montague PR (2007) Neural signature of fictive learning signals in a sequential investment task. *Proc Natl Acad Sci USA* 104:9493–9498.
6. Friston J, *et al.* (1995) Statistical parametric maps in functional imaging: A general linear approach. *Hum Brain Mapp* 2:189–210.
7. Frith U, Frith CD (2003) Development and neurophysiology of mentalizing. *Philos Trans R Soc London Ser B* 358:459–473.

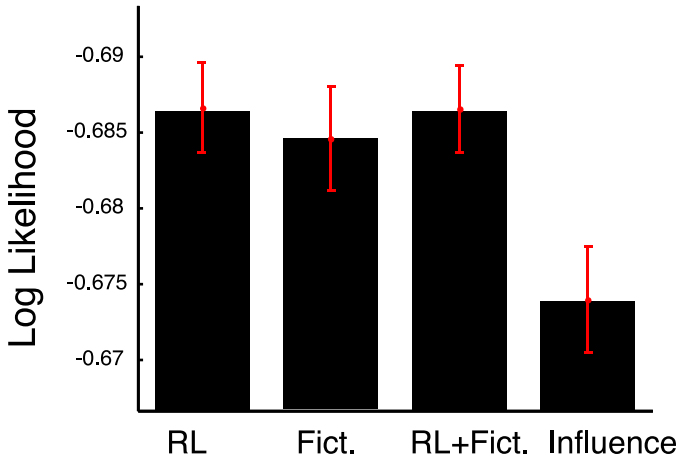

Fig. S1. Out-of-sample model log likelihood. The out-of-sample model log likelihood controls for models having different number of free parameters when fitting to behavioral data. Models were trained with the first 70 trials for each subject and then tested on the last 30 trials to obtain an out-of-sample log likelihood. The influence model accounts for subjects' behavior the best, with an out-of-sample log likelihood of 0.674 ± 0.004, followed by the fictitious play model with 0.685 ± 0.003 and the RL model with 0.687 ± 0.003.

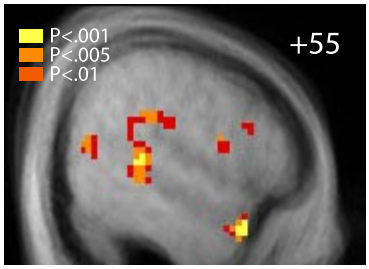

Fig. S2. Model comparisons with respect to the processing of expected reward signals in the brain. The influence model expected reward signals that are not explained by (orthogonal to) the RL model expected reward signals also activate the right STS, including the right temporal pole at $P < 0.001$ uncorrected.

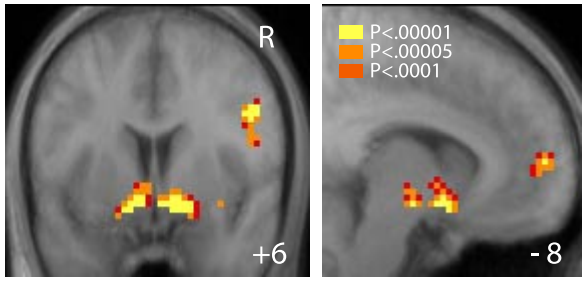

Fig. S3. Prediction error signals. The prediction error signals generated by the influence model were correlated with activity in the ventral striatum bilaterally (9, 6, −18 mm, $z = 4.97$; −9, 9, −18 mm, $z = 4.73$, both $P < 0.05$ whole-brain corrected), mPFC (−9, 57, 6 mm, $z = 4.35$), and paracingulate cortex (12, 36, 18 mm, $z = 4.62$). This lends support to the suggestion that mPFC is not only involved in calculating expected reward signals derived from inference of the opponent's game strategy (Fig. 2*A*), but is also involved in the update of the inferred opponent's strategy through prediction errors (this figure) and influence updates (Fig. 3*B*).

Table S1. Fitted parameters

Model	Parameter	Employer	Employee
RL	β	0.011	0.025
	η	0.16	0.11
Fictitious play	β	0.036	0.056
	η	0.22	0.10
Influence model	β	0.090	0.059
	η	0.21	0.038
	κ	0.0011	0.043